

الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

المدرسة العليا للإعلام الآلي ٨ ماي ١٩٤٥ - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Graduation Thesis

To obtain the diploma of **Engineer**
Field of Study: **Computer Science**
Specialization: **IASD**

Theme

**From Text Prompts to Geospatial Insights: A Unified VLM-Based Pipeline
for Multi-Species Tree Mapping and Health Assessment**

Presented by
Meflah Yousra

Defended on: **July 6, 2026**
In front of the jury composed of

Mr. [Jury Member Name]
Mr. Larabi Mohammed ElAmine
Mr. Iftene Meziane
Mr. Chaib Souleyman
Mr. [Jury Member Name]

President of the Jury
Thesis Supervisor
Co-Supervisor
Co-Supervisor
Examiner

Academic Year: 2025/2026

Abstract

Monitoring individual trees at scale is important for carbon accounting, precision agriculture, and food security, but traditional field surveys are slow and hard to scale. Deep learning has improved tree detection from remote sensing imagery, yet most systems only locate trees without assessing their health or species. This thesis presents GeoRS-CLIP, a vision-language system for multi-species tree detection and health assessment from remote sensing images. The system has three stages: (1) an open-vocabulary detector based on CLIP, conditioned on geographic and spectral information through a module called GeoFiLM, (2) zero-shot instance segmentation using SAM2 to extract tree crown masks from detection boxes, and (3) a vision-language reasoning step using Qwen3-VL that generates interpretable health explanations from RGB crops and spectral indices.

GeoFiLM is the main contribution: it injects spectral indices (NDVI, ExG) and geographic metadata into a frozen CLIP ViT-B/16 backbone using lightweight feature modulation, adding only 445K parameters (0.3% of the backbone). On a dataset of 34,000+ images covering urban trees, date palms, and olive trees, GeoFiLM improves detection by +20.3% mAP@0.5 and +13.5% F1-score over the baseline. A cross-attention classification head further reduces species misclassification from 23% to 8%, and the final multi-species model reaches 0.727 mAP@0.5. SAM2 converts detection boxes into tree crown masks without any pixel-level training. Qwen3-VL then generates short health reports from each detected tree using its RGB appearance and spectral values, providing interpretable outputs for agronomic use.

A qualitative zero-shot case study on Algerian date palm oases (Biskra) and olive orchards (Tizi Ouzou) demonstrates the system’s applicability to new regions using satellite imagery, without retraining. While quantitative generalization metrics are not reported due to the absence of labelled ground-truth for these regions, the qualitative results show coherent crown delineation and plausible health reasoning. For health assessment, although no labelled health ground-truth exists for statistical validation, the reasoning module provides an auditable chain-of-thought that domain experts can review. Together, these results demonstrate the potential of the approach as a low-cost decision-support tool for vegetation monitoring in data-scarce environments.

Keywords: Remote Sensing, Tree Detection, Vision-Language Models, Open-Vocabulary Detection, Instance Segmentation, Tree Health Assessment, Precision Agriculture

Résumé

Le suivi des arbres individuels à grande échelle est important pour la comptabilisation du carbone, l'agriculture de précision et la sécurité alimentaire, mais les relevés de terrain traditionnels sont lents et difficiles à généraliser. L'apprentissage profond a amélioré la détection des arbres à partir d'images de télédétection, mais la plupart des systèmes se limitent à la localisation sans évaluer leur santé ou leur espèce.

Cette thèse présente GeoRS-CLIP, un système vision-langage pour la détection multi-espèces et l'évaluation de la santé des arbres à partir d'images de télédétection. Le système comporte trois étapes : (1) un détecteur à vocabulaire ouvert basé sur CLIP, conditionné par des informations géographiques et spectrales via GeoFiLM, (2) une segmentation d'instance zero-shot utilisant SAM2 pour extraire les masques de couronne, et (3) une étape de raisonnement vision-langage utilisant Qwen3-VL qui génère des explications interprétables sur la santé des arbres.

GeoFiLM constitue la contribution principale : il injecte des indices spectraux (NDVI, ExG) et des métadonnées géographiques dans un backbone CLIP ViT-B/16 gelé, en ajoutant seulement 445K paramètres (0,3% du backbone). Sur un jeu de données de plus de 34 000 images couvrant les arbres urbains, les palmiers dattiers et les oliviers, GeoFiLM améliore la détection de +20,3% en mAP@0.5 et +13,5% en F1-score. Une tête de classification à attention croisée réduit les erreurs de classification d'espèces de 23% à 8%, et le modèle multi-espèces final atteint 0,727 de mAP@0.5.

SAM2 convertit les boîtes de détection en masques de couronne sans entraînement au niveau pixel. Qwen3-VL génère ensuite des rapports de santé interprétables pour chaque arbre détecté.

Une étude de cas qualitative zero-shot sur les oasis de palmiers dattiers algériennes (Biskra) et les oliveraies (Tizi Ouzou) démontre l'applicabilité du système à de nouvelles régions sans réentraînement. Bien que les métriques quantitatives ne soient pas rapportées en raison de l'absence de vérité terrain annotée, les résultats qualitatifs montrent une délimitation cohérente des couronnes et un raisonnement sanitaire plausible. Pour l'évaluation de la santé, le module de raisonnement fournit une chaîne de raisonnement vérifiable que les experts peuvent examiner. Ces résultats démontrent le potentiel de l'approche comme outil d'aide à la décision à faible coût pour le suivi de la végétation dans les environnements pauvres en données.

Mots-clés : Télédétection, Détection d'Arbres, Modèles Vision-Langage, Détection à Vocabulaire Ouvert, Segmentation d'Instance, Évaluation de la Santé des Arbres, Agriculture de Précision

الملخص

يُعدّ رصد الأشجار الفردية على نطاق واسع أمراً مهماً لحساب الكربون، والزراعة الدقيقة، والأمن الغذائي، إلا أن المسوحات الميدانية التقليدية بطيئة وصعبة التوسّع. لقد حسّن التعلم العميق من كشف الأشجار انطلاقاً من صور الاستشعار عن بُعد، غير أن معظم الأنظمة تقتصر على تحديد الموقع دون تقييم صحة الشجرة أو نوعها.

تقدّم هذه الأطروحة نظام GeoRS-CLIP، وهو نظام رؤية-لغة للكشف متعدد الأنواع وتقييم صحة الأشجار. يتكوّن النظام من ثلاث مراحل: (1) كاشف ذو مفردات مفتوحة مبني على CLIP، مشروط بمعلومات جغرافية وطيفية عبر وحدة GeoFiLM، (2) تجزئة كائنات بدون تدريب باستخدام SAM2 لاستخراج أقنعة تاج الشجرة، و(3) استدلال رؤية-لغة باستخدام Qwen3-VL يُولّد تفسيرات صحية قابلة للتفسير.

تُعدّ وحدة GeoFiLM المساهمة الرئيسية: فهي تُدخل مؤشرات طيفية وبيانات جغرافية إلى نموذج CLIP المُجمّد، بإضافة 445 ألف معامل فقط (0.3% من النموذج الأساسي). وعلى مجموعة بيانات تضم أكثر من 34000 صورة، حسّنت GeoFiLM دقة الكشف بنسبة +3.20% في $mAP@0.5$ و+5.13% في $F1$ كما خفّضت رأس تصنيف قائم على الانتباه المتقاطع نسبةً لخطأ في تصنيف الأنواع من 23% إلى 8%، ووصل النموذج النهائي إلى دقة 727.0 في $mAP@0.5$.

يحوّل SAM2 صناديق الكشف إلى أقنعة تاج الشجرة، ثم يُولّد Qwen3-VL تقارير صحية قصيرة لكل شجرة مكتشفة.

وتُظهر دراسة حالة نوعية على واحات نخيل التمر ببسكرة وبساتين الزيتون بتيزي وزو بالجزائر، قابلية تطبيق النظام على مناطق جديدة دون إعادة تدريب. ورغم أن المقاييس الكمية غير مُبلّغة لعدم وجود بيانات مرجعية، فإن النتائج النوعية تُظهر تحديداً متماسكاً واستدلالاً صحيحاً معقولاً. وتُبرز هذه النتائج إمكانات هذا النهج كأداة دعم قرار منخفضة التكلفة لرصد الغطاء النباتي في البيئات الفقيرة بالبيانات.

الكلمات المفتاحية: الاستشعار عن بُعد، كشف الأشجار، نماذج الرؤية واللغة، الكشف ذو المفردات المفتوحة، تجزئة الكائنات، تقييم صحة الأشجار، الزراعة الدقيقة